ES 335: Machine Learning

Quiz — Trees, Metrics & Ensembles (2023 archive) — ANSWER KEY

Duration: 45 minutes Maximum marks: 9

SET A

Instructions

- Write your name and roll number before starting.
- This paper has 8 questions in 3 parts, for a total of 9 marks.

MCQ Answer Grid

Q	1	2
Ans	**D**	**A**

Part A — Multiple Choice

2 marks

Q1. For the Tennis example the maximum entropy is 1.0 bit. What is the maximum entropy of an ImageNet-style classification problem with 1024 classes? *[1 mark]*

(A) 1.0 bit
(B) 512 bits
(C) 1024 bits
(D) 10 bits ✓

> **Explanation.** Entropy is maximized by the uniform distribution: $\log_2 1024 = 10$ bits.

Q2. In the lectures we saw that `np.std(x)` and `pd.Series(x).std()` give different values on the same data. Why? *[1 mark]*

(A) NumPy defaults to the population estimate (divides by N, `ddof=0`) while pandas defaults to the sample estimate (divides by $N-1$, `ddof=1`) ✓
(B) They use different definitions of the mean
(C) pandas rounds the result to six decimal places by default
(D) NumPy accumulates in `float32`, losing precision on large arrays

Part B — Short Answers

4.5 marks

Answer in the space provided; show your working.

Q3. Create an example ground truth and prediction where the mean absolute error is 100 and the mean error is 0. *[0.5 marks]*

> **Model answer.** Any symmetric errors, e.g. $y = (0, 0)$, $\hat{y} = (100, -100)$: ME $= 0$, MAE $= 100$.

Q4. Given the following dataset, which attribute would the decision-tree algorithm split on first? Why? *[1 mark]*

Sample #	Radius	Weight	Color	Quality

1	1	1	1	Good
2	1	1	2	Good
3	1	2	1	Bad
4	1	2	2	Bad
5	2	1	1	Good
6	2	2	2	Good

Model answer. Weight. $H(\text{Quality}) = H(2/6) \approx 0.918$. Splitting on Weight: weight 1 → all Good ($H = 0$); weight 2 → (1 Good, 2 Bad), $H \approx 0.918$; weighted ≈ 0.459, so IG ≈ 0.459 — higher than Radius (IG ≈ 0.25) and Color (IG $= 0$).

Rubric. +0.5 correct attribute; +0.5 information-gain argument.

Q5. Quoting Wikipedia: *[2 marks]*

> Pre-pruning procedures prevent a complete induction of the training set by replacing a stop criterion in the induction algorithm (e.g. maximum tree depth or information gain > minGain).

Create a decision tree for the classification problem below, and explain why pre-pruning using information gain can be limiting.

x_1	x_2	y
0	0	0
0	1	1
1	0	1
1	1	0

Model answer. This is XOR: splitting on either x_1 or x_2 alone leaves both children at 50/50, so IG $= 0$ at the root — gain-based pre-pruning halts immediately. Yet the depth-2 tree (split x_1, then x_2 in each branch) classifies perfectly: gain can be zero at one level and large at the next.

Rubric. +1 a correct depth-2 tree (split on x_1, then x_2); +1 argument: both root splits have zero gain, so gain-based pre-pruning stops before the informative second level.

Q6. Create one confusion matrix for 100 total examples where the precision is 0.8 and the recall is 0.5. *[1 mark]*

Model answer. Take TP $= 40$: precision $0.8 \Rightarrow$ FP $= 10$; recall $0.5 \Rightarrow$ FN $= 40$; then TN $= 10$. Total $= 100$.

Rubric. +0.5 consistent TP/FP from precision; +0.5 FN from recall and a total of 100.

Part C — Ensembles *2.5 marks*

Q7. In bootstrap sampling we sample N times with replacement from an original dataset of N distinct elements (e.g. for $N = 8$, a bootstrap sample may be $\{8, 8, 3, 4, 5, 1, 8, 5\}$, which has 5 unique elements). Show that on average a bootstrap sample contains $\approx 63.2\%$ of N unique elements. *[1.5 marks]*

Model answer. An element is missed by one draw w.p. $1 - 1/N$, by all N draws w.p. $(1 - 1/N)^N$. Expected unique fraction $= 1 - (1 - 1/N)^N \to 1 - 1/e \approx 0.632$ as N grows.

Rubric. $+0.5$ miss probability $(1 - 1/N)^N$; $+0.5$ expected unique count $N(1 - (1 - 1/N)^N)$; $+0.5$ limit $1 - 1/e \approx 0.632$.

Q8. Assume K ensemble members, each with the same error probability $p < 0.5$. The binomial argument says the ensemble error shrinks as K grows — yet empirically, adding members may not always reduce the error. Why? *[1 mark]*

Model answer. The binomial argument assumes **independent** errors. Members trained on similar data make correlated mistakes, so the effective ensemble size saturates and errors shared by many members never get voted away.

— End of paper —